DIGITAL ARCHIVING AT NASA GODDARD SPACE FLIGHT CENTER

Part 2:

The Digital Archiving Plan: Moving Toward an Infrastructure

NASA GSFC Library

Draft 2
September 17, 2002
Revised Final Draft
December 20, 2002

Table of Contents

1.0 GOALS

The goal of Part 2 of the GSFC Digital Archiving Plan is to provide a framework and a roadmap for the development of digital archiving procedures in support of GSFC's broader knowledge management objectives. In addition, it discusses the scope of a proposed implementation, sets out the activities that must be done to provide a more comprehensive, operational plan, and estimates system and human resources needed to further develop the framework into an operational system.

2.0 SCOPE OF THE PLAN

Based on the results presented in Part 1 of this report, GSFC has a variety of sources and information types that may be of interest to scientists and engineers in the near term and into the future. Based on the analysis of the key engineering and science web sites used in the pilot, the project documentation and web-based information covers a broad span of the anticipated object types and formats, from text to video to data.

In addition to the scope of the information to be collected, this plan is constrained by the degree to which direction can be given to, and subsequent compliance achieved by non-Library organizations. Rather than present a large-scale approach that requires extensive coordination and resources across the organization, this approach presents a number of alternatives that, in combination, can be used to gather the objects of permanent value and to work toward a more comprehensive approach in the future.

3.0 GUIDING PRINCIPLES

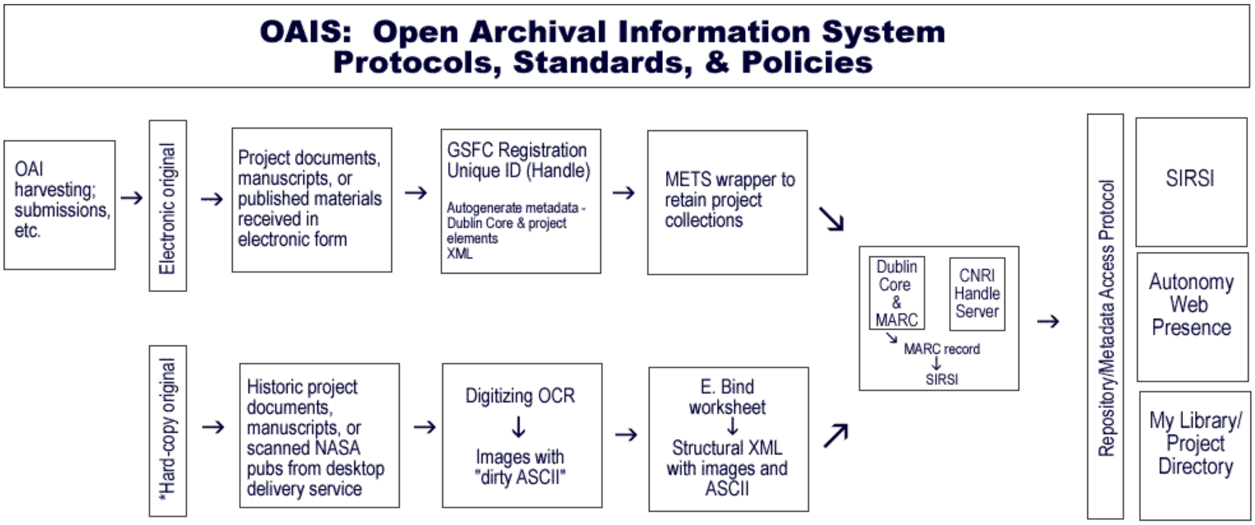
The following served as guiding principles for the development of the plan.

- The Open Archival Information System (OAIS) Reference Model will serve as the overarching conceptual framework for the GSFC digital archive.
- The framework and the plan should address all object types of interest to GSFC, even if not all the objects can be archived at the outset.
- The plan should be flexible and extensible to accommodate as yet unknown digital objects and topics of interest to the GSFC community.
- The system will be standards-based as much as possible, and standards will be extended to handle GSFC-specific information where needed.
- GSFC will work with related communities of practice to develop standards and best practices where none exist.
- The plan should take advantage of other initiatives at GSFC, particularly those related to knowledge management.
- A centralized trusted third-party model will be used, with the GSFC Library acquiring, via spidering, harvesting and deposit, the materials for completed projects and other digital objects.

- The ultimate repository for materials and the ultimate place for their stewardship and long term preservation will be a centralized archive run by the GSFC Library.
- A staged implementation plan will be used.

4.0 Conceptual Framework

The conceptual framework is based on the OAIS Reference Model (see Part 1, Section 4.1, Fig. 1.). Its goal is to use current standards and best practices, as well as the backbone and expertise of the GSFC Library to support long-term digital preservation and access. The more detailed infrastructure model is built on the model developed by the National Agricultural Library to support the US Department of Agriculture's project to preserve its digital publications (see Part 1, Section 4.1, Fig. 2).


OAIS: Open Archival Information System
Protocols, Standards, & Policies
OAI harvesting; submissions, etc.
Electronic original
Project documents, manuscripts, or published materials received in electronic form
GSFC Registration Unique ID (Handle)
Autogenerate metadata - Dublin Core & project elements XML
METS wrapper to retain project collections
Dublin Core & MARC
CNRI Handle Server
MARC record
SIRSI
*Hard-copy original
Historic project documents, manuscripts, or scanned NASA pubs from desktop delivery service
Digitizing OCR
Images with "dirty ASCII"
E. Bind worksheet
Structural XML with images and ASCII
Repository/Metadata Access Protocol
SIRSI
Autonomy Web Presence
My Library/ Project Directory

*This is not expected to be a major focus of the prototype. However, it is important to consider the possibility of non-digital materials, both current and legacy. Digitization may be needed for unusual formats.

Figure 1. Conceptual framework

5.0 Framework for Guidelines

When archiving digital objects there are several key areas that must be addressed. Since the scope of this project does not include the development of the specific guidelines for archiving, these issues are raised along with recommendations for how to develop the specific guidelines. The recommendations are informed by the two pilot projects that were conducted and by projects conducted by other groups both nationally and internationally as discussed in Part 1 of this report. The guidelines are put in the context of the OAIS and the conceptual framework for handling electronic originals described above.

5.1 Creation by Producer

Creation is the act of producing the information product. The producer may be a human author or originator, or a piece of equipment such as a sensing device, satellite or laboratory instrument. Long-term preservation starts at the creation stage, since even in rigorously controlled situations, information can be lost if the originator does not realize the importance of its retention. Practices used when a digital object is created have an impact on the ease with which the object can be digitally preserved. In the case of the GSFC, most of the producers with responsibility for the creation of the information are internal to the center or are funded by center programs.

In this “corporate” setting, the creators can support the process by using more standardized applications for creation of the input. They can also be encouraged through policies and other incentives to contribute to the archive and to provide metadata. **GSFC should develop templates and applications that would support the creation of various object types, particularly the documents related to projects. The analysis shows that most files are ASCII text, Word, Excel or pdf. Analysis of the guidelines provided for projects shows that there are already standard formats in place for the creation of various project reports and documents.** Converting these formats or identifying cases where they have already been converted to templates would make the creation of these documents easier and provide a method for the creator to create metadata in the regular workflow. The best practice would be to create the metadata at the object creation stage, or to create the metadata in stages, with the metadata provided at creation augmented by additional elements as needed when the materials reach the archive. **In addition, key information, such as lessons learned that are embedded in text could be tagged using in-line XML for extraction or online retrieval.**

The creators can provide an indication of the anticipated long-term value of the information. **The National Library of Medicine’s Permanence Rating System should be used to express the long-term value of the material.** (The Permanence Rating System developed by the National Library of Medicine is included in Part 1, Appendix C.) The creator could apply these permanence ratings. In lieu of other assessment factors, the creator’s estimate can serve as an indication of the value that will be placed on the object by people within the same discipline or area of research in the future. This rating would not take the place of formal retention schedules for objects that qualify as electronic records.

Other ratings may be needed to accommodate the needs at GSFC, but they should be developed in the context of the NLM model, which has already been vetted and is being considered as an ad hoc standard both nationally and internationally. The Permanence Rating System for GSFC should be developed by the Library along with the Steering Committee.

5.2 Ingest

Ingest is the stage in which the created object is “incorporated” physically or virtually into the archive. The object must be known to the archive administration. In order to ingest objects, the Archive must establish collection policies and methodologies for gathering the content.

5.2.1 Collection Policies

The collection policies answer questions related to selecting what to archive, determining extent, archiving links, and refreshing site contents.

5.2.2 Selecting What to Archive

In order to develop selection guidelines, **GSFC will need to achieve some level of consensus about the goal of preserving digital materials.** There are several reasons for preservation including retaining the history of NASA, providing for disaster recovery and backup, addressing future legal and auditing issues, and providing information needed to advance GSFC's engineering and scientific missions in the wake of human capital concerns.

Once the reason or reasons for the archive are identified, it will be easier to determine what objects should be kept in the GSFC Archive and for how long. This decision will then be reflected in written guidelines and in the Permanence Rating System described in section 5.1 above.

As an initial scope, the team recommends continued work on the videos for colloquia and other lecture series; the current project web sites included in the Project Directory maintained by the GSFC Library and any new projects identified; and web pages with relevant content from the engineering and scientific codes.

The team recommends that the initial implementation should not include data sets, video or audio clips, software, etc. that are linked from the web pages. These files can be identified by format type (jpg, mpg, etc.). However, the collection of limited object types should not preclude the coordination with data and images archives that already exist at GSFC. Eventually, these object types will be brought into the future pilots and into the operational system, but it is important to address the complex issues of digital archiving and preservation in a well-planned manner.

The selection guidelines should be informed by input from the stakeholder groups, in particular the potential users of the archive. Some input has already been received as part of the selection criteria for the CIO Pilot Project (see Part 2, Appendix A). The consultant recommends that the Steering Committee establish a task group to specifically address the issues of selection. It is necessary to balance the potential needs of the users against the constraints such as size, retrieval speed, application software, format and intellectual property concerns.

5.2.3 Determining Extent

Directly connected to the question of selection is the issue of extent. What is the extent or the boundary of a digital work? This issue arises particularly with complex web sites. **Generally, for purposes of preservation and cataloging, preference is given to breaking down large sites into component titles and selecting those that meet the general guidelines. This also allows metadata to be provided for a set of pages that are intellectually cohesive. However, sometimes the components of larger sites do not stand well on their own, in terms**

of the intellectual content. In these cases, the sites should be selected for archiving as an entire entity.

The analysis of the project pages showed that significant scientific content occurs between levels 3 and 10.   In many recent sites, the homepage is introductory and is not particularly content rich (see Part 1, Appendix I; RHESSI Home Page, Planet B or Hubble Space Telescop). However, the pilot also showed that especially with project sites the information is so extensive that to save the information down 10 levels will produce terabytes of information and take weeks of gathering. It also tends to spider the content from .com sites, which may be copyrighted.

5.2.4 Archiving Linked Objects

The extensive use of hypertext links to other digital objects in electronic publications raises the question of whether the text of these links should be archived along with the source item. **The consultant recommends that minimally, the text to .com links, which are probably copyright should not be captured.**

However, even outside the .com links there are issues that need more discussion. Links could be retained when they are on the same originating server or in the NASA or another .gov domain. Linked text could be archived only if it meets the selection guidelines. The links (without the text) could be retained as live, but the system could underscore the fact that later access may find them broken. Alternatively the URLs could be retained but the text not captured unless the site is on the originating server or in the NASA or another .gov domain. A compromise may be to establish a system whereby links to project partners are maintained but others are not. **This requires more consideration as specific guidelines are developed. The guidelines should take into consideration current NASA web guidelines and general best practices as identified by federal directives and standards bodies such as the World Wide Web Consortium.**

Ultimately, GSFC should develop a Handle resolver service that could be used by GSFC, its partners and contractors. The resolver is described in Part 1 section 4.3.6 and Part 2, section 5.5.1. Such a system would allow material to be moved and reorganized in its physical location while keeping the persistent identification provided by the Handle. This will be particularly important if the federated archive approach is used. See Part 1, section 4.5.3.

5.2.5 Intellectual Property Concerns

Despite the fact that much of the GSFC information is in the public domain, there are intellectual property concerns that must be addressed in the archive. The guidelines should address the issues of security and proprietary rights using a system similar to that used for distribution limitations on technical reports and other documents that must be controlled. GSFC should investigate the use of the <indecs> system that is being implemented by the DOI Foundation (www.indecs.org) While this is primarily for management of copyrighted, journal material, the DOI Foundation has expressed an interest in extending such a scheme to manage government information.

Unfortunately, unless all .coms and .edus are considered off limits to a spider, it will be necessary to evaluate the intellectual property rights associated with linked pages in these domains. This would need to be done on a case-by-case basis either as part of the initial page development or at the point of ingest. There may be some scripts that could be written to reflect selection logic, but it isn't likely that this would be 100% accurate. Once the initial archive has been created for projects that have already been completed, the identification of property rights should be part of the management of the project library and the turnover of the content to the Archive.

Possible changes to the default data clauses in contracts and grants should be considered in order to make a more viable archiving regime. Even if access only within GSFC or within NASA were allowed for the linked sites this would be preferred to wide-scale site-by-site evaluation for copyrighted and proprietary material. Project management might also consider changing contractor practices so that project information that is hosted by a non-GSFC entity would be provided to the Archive, along with an indication of the appropriate distribution limitations, when the project is completed.

5.3 Gathering Content

5.3.1 Spiders, Harvesters, and Deposit

There are three basic approaches to the gathering of web-based resources – spiders (or robots), harvesters, or deposit. In the first approach, the burden of collection is on the archive, which sets up schedules and parameters for automatic copying of digital objects. In the second approach, the archive targets the specific URLs and establishes mechanisms for copying only certain pages or metadata from those pages. In the third approach, the Project Libraries or other originators would submit or deposit resources to the Archive. This is generally done via ftp because of the size of the files.

The team suggests that all three approaches are valid. When the archive is first implemented, spidering will be needed to capture legacy information. Otherwise, the gathering of this information would be too resource intensive. Priority can be given to various types of docuhhments based on mime type. Conferences, usenet groups, ftp archives, and databases could be eliminated based on format types.

However, once the legacy information has been gathered, it will be possible to target the new projects that come into the Project Directory and to monitor science and engineering code pages for changes. The harvesting approach could then be used, allowing for more precise selection of the extent of pages and review of the metadata content.

Ultimately, the GSFC culture of knowledge management may move toward a more depository approach. This would be in line with the approaches, incentives and mechanisms discussed in the Knowledge Management Strategic Plan and the EOS Lessons Learned project. This has the benefit of saving Library resources and increasing the community's stake in the archive and its continuity. However, this approach may result in difficulties based on inconsistent content if best practices and standards are not established. Spidering and harvesting should always be available back up methods to ensure content is being gathered.

5.3.2 Minimizing the Impact on Target Sites

As already noted in Part 1, it is important to consider the impact of the gathering effort on the target sites. Time limits between visits should be established. Most of the efforts should be scheduled for nights and weekends. The best approaches should be worked out through consensus with the GSFC webmasters group, and the schedule should be posted and widely announced. The flexibility that is ultimately needed may be at the URL level, which may require a different software program or the integration of the program with a scheduler that is a database application, which can be easily modified by the digital archivist.

5.3.3 Recapturing Archived Sites

In cases where the archiving is taking place while changes or updates may still be occurring to the digital object, there is a need to consider recapturing the contents of previously archived sites. A balance must be struck between the completeness and currency of the archive and the burden on the system resources. Obviously, the burden of recapturing the contents increases as the number of items stored in the archive increases.

Project sites and the information from the pages of the engineering and science codes will change over time. Further analysis of these object types will help the Library to determine a gathering schedule by type or code so that the automatic harvesting program can be limited. The options might include on/off, weekly, monthly, quarterly, every six months, every nine months, or annually.

The selection is dependent on the degree of change expected and the overall stability of the site. It is assumed that users will access the “current” pages for active projects. Once a project is completed or becomes inactive, there is little activity/change to the site. However, the team found some instances where changes did occur. However, a default schedule should be established to ensure that information is relatively up-to-date even if the automatic selective mechanisms fail. In the analysis of the pilot test on web sites, a schedule of once per quarter seems most efficient as the default value.

The degree of change is also a factor in the Permanence Rating System, allowing the rating assigned by the creators to be used automatically to determine the schedule for recapturing the contents.

5.4 Data Management

Once the archive has acquired the digital object, it is necessary to manage it through identification and cataloging. Both identification and cataloging allow the archiving organization to manage the digital objects over time. Identification provides a unique key for finding the object and linking that object to other related objects. Cataloging in the form of metadata supports organization, access and preservation. Cataloging and identification practices may differ based on what is being archived and the resources available for managing the Archive.

5.4.1 Maintaining Collections

Documents and other objects for each particular project are essentially collections within a large digital library collection made up of all GSFC projects and ultimately all GSFC material. Within each project, there are numerous project-related objects that can be of any format or object type.

The consultant recommends that the archiving system support the "collective nature" of the project documents both in support of GSFC and the broader interagency and agency level. What does this mean for the system?

First it means that metadata must be developed and stored for even the highest collection level at GSFC. The archival repository at GSFC must have metadata that minimally reflects the best practices for collection description that can be identified. Ultimately, it would be valuable to have consensus on this metadata across the NASA centers and among other government agencies and non-government partners.

In addition to metadata, the system must support the re-creation of the project collections. In addition to descriptive metadata, structural metadata is important. In support of this concept, this project included an analysis of the METS standard, which is discussed in detail in Part 1, section 5.3 and Appendix E.

5.4.2 Metadata

All archives use some form of metadata for description, reuse, administration, and preservation of the archived object. There are issues related to how the metadata is created, the level at which metadata is applied, the metadata standards and content rules that are used, and where the metadata is stored.

5.4.2.1 Metadata Creation

Because of limited resources for the Archive, the majority of the metadata must be created automatically or at the point of creation. Section 5.1 above discusses the use of templates to aid in the creation of metadata by the producers of the information. Some metadata, such as file type, document type, date, etc. can be created automatically or candidate information can be created based on rules and assumptions. For Dublin Core metadata, the dc.dot metadata generator tool can be used to create candidate metatags. After creation of the template, the XML generator can be used to create an XML stream for archiving the metadata or importing to a database for searching. It will be necessary to have some metadata review at the point at which the material is ingested into the archive. This will be particularly important if the automatic generation is used as described.

The use of a Dublin Core generator and an XML generator are being piloted in the product system pilot currently underway. As described in Part 1, Section 5.2.6, there are problems with the reliability of this automated approach. If resources can be identified, the quality would be greatly improved by having project librarians support the Archive through review and enhancement of the metadata creation. Alternatively, this could be done by the Archive. However, once support for metadata creation is provided to creators of the original project documents and web sites, the need for this work (outside the routine flow) should be minimized.

5.4.2.2 Application of Metadata to Various Levels

The level at which metadata is applied depends on the object type, the document type and the permanence rating. However, in general, metadata must be provided in order to keep the collection together intellectually and structurally (as outlined in Section 5.4.1 above) and to match the level of extent of the item (see section 5.2.3).

5.4.2.3 Metadata Standards

Based on the communities of practice involved, the rationale for the metadata and its potential uses, and the standards that can be implemented, metadata standards must be selected. In the case of GSFC materials, a wide range of object types are involved and therefore, multiple metadata standards are necessary.

Within the framework of METS and the OAIS, the following metadata standards will be used and integrated.

5.4.2.3.1 Descriptive Metadata

For descriptive metadata the Dublin Core will be used (see Part 1, Appendix A). This provides a basic catalog record for source, creator, dates, subject and title. There are 15 elements, all of which are considered optional by the standard. However, GSFC would need to make some of these elements mandatory. For example, the creator, date and title.

5.4.2.3.2 Geospatial Metadata

Some of the information to be archived will have geospatial reference components, for example some of the Global Change Master Directory information. In this case, the Federal Geographic Data Committee Content Standard applies. These elements have been mapped to Dublin Core or they can be layered to allow for an extension to represent this type of information.

5.4.2.3.3 Metadata for Preservation

RLG and OCLC have developed a draft standard for metadata for preservation (see Part 1, Appendix B). While much of this standard is text based, it can be used for the majority of the types of material that GSFC would initially be interested in.

5.4.2.3.4 Metadata for Specific Science Specialties

The individual communities of practice should identify the metadata that they might find of value. Some of this can be investigated for example with the IEEE and IEE. There is already a Biological Profile for the FGDC metadata standard, which can be used for some of the environmental and life science information at GSFC. Some groups such as the Global Change Master Directory already have sets available that could be used.

5.4.2.3.5 Metadata for Specific Object Types

The trial standard from NISO and AIIM can be used as part of the metadata set when the format type indicates a digital still image. This standard should be used with systems such as NIX that are already being improved by the Library.

There is no standard metadata for videos. However, efforts are underway among various groups, including the Corporation for Public Broadcasting, which should be followed for later incorporation into the GSFC plans.

5.4.2.3.6 Project Metadata

In a recent project to automatically categorize project documents, the user testing of the system highlighted the fact that there were taxonomies other than subject that users would like to use to retrieve or limit searches of such documents. Some of these are related to mission or project document type. The upcoming analysis of metadata across the EOS Project Libraries for the EOS Lessons Learned Prototype will help to further inform the project related metadata elements that should be included.

5.4.2.4 Content Rules

While some metadata schemes dictate rules for completing the content of the tags, most do not. The standard can reference enumerated lists, coded values and local lists for both input and validation. **GSFC must determine the elements that need to be controlled and then determine the specific controlled lists that should be used.**

5.4.2.5 Where the Metadata is Stored

With some objects, such as web pages, the metadata can be stored with the object. In other cases, it is more difficult to do this and linked, external metadata is the more appropriate approach.

For purposes of preservation, the team suggests that the metadata be stored with the original object. However, for consistency and efficient resource discovery, the metadata from all types of objects should be stored in one database, external from the objects. The creation of a database or clearinghouse of metadata related to the GSFC Archive can be facilitated by writing or identifying scripts that will turn the metadata into XML for import into a structured database. It is also recommended that when possible the metadata should be stored embedded with the item, for example as HTML or XHTML metatags.

5.5 Archival Storage

Storage is often treated as a passive stage in the life cycle, but storage media and formats have changed with legacy information being lost forever. The most common solution to this problem of changing storage media is migration to new storage systems. This is expensive, and there is always concern about the loss of data or problems with the quality when a transfer is made. Check algorithms are extremely important when this approach is used.

5.5.1 Persistent Identification

For those archives that do not copy the digital material immediately into the archive, the movement of material from server to server or from directory to directory on the network, resulting in a change in the URL, is problematic. Using the server address as the location identifier may result in a lack of persistence over time both for the source object and any linked objects.

In order to bypass any problems with location, the GSFC archiving system must determine a scheme for object identification. The Handle is recommended since it can provide for multiple resolutions (see Part 1, Section 4.3.6). An alternative number such as the AAS ID could also be included in the file for those documents that are also covered in the AAS database.

5.5.2 Transformation vs. Native Formats

A key preservation issue is the format in which the archival version should be stored. Transformation is the process of converting the native format to a standard format. The final decision will depend on the types of objects, the native formats, the availability of general archival formats that are likely to persevere into the future, and the ability to use these archival formats to render the object in a state that is close to its native format.

5.6 Preservation Planning

Preservation keeps both the content and the look and feel of the digital object. The definition of "long-term" will vary depending on the user community. Depending on the particular technologies and subject disciplines involved, the cycle for hardware/software migration may be 2-10 years.

5.6.1 Hardware and Software Migration

A critical part of preservation planning is making decisions about hardware and software migration. Plans are less rigorous for migrating to new hardware and applications software than for storage media. New releases of databases, spreadsheets, and word processors can be expected at least every two to three years, with patches and minor updates released more often. While software vendors generally provide migration strategies or upward compatibility for some generations of their products, this may not be true beyond one or two generations. Migration is not guaranteed to work for all data types, and it becomes particularly unreliable if the information product has used sophisticated software features. There is generally no backward compatibility, and if it is possible, there is certainly loss of integrity in the result.

In order to guard against major hardware/software migration issues, the organizations try to procure mainstream commercial technologies. Unfortunately, this level of standardization and ease of migration is not as readily available among technologies used in specialized fields such as those at GSFC where niche systems are required because of the interfaces to instrumentation and the volume of data to be stored and manipulated.

The best practice for the foreseeable future will be migration to newhardware and software platforms; emulation will begin to be used if and when the hardware and software industries begin to endorse it.

5.6.2 Preservation of the Look and Feel

Preserving the look and feel of the digital object is more difficult than preserving the content. The degree to which the look and feel must be retained and later made accessible depends on the type of object and the way the user community will reuse the content.

Pdf, TIFF IV and ASCII are common formats for archival storage, with the latter being relatively standardized. Pdf has been used to maintain the look and feel across platforms and devices, and this is also a very prevalent format for content in project libraries and as linked documents to GSFC web sites. However, concerns have been raised about the proprietary nature of this format. As pdf has become more prevalent, these concerns are being erased in part by the force of the marketplace and the fact that organizations such as NARA are working on preservation mechanisms for pdf.

5.6.3 Validation and Security

As part of the ingest and dissemination processes, the Archive should implement checksum routines and other types of validation to ensure that bit loss has not resulted from the moving and processing the bit streams. Security can be addressed through the use of the regular GSFC firewalls or by implementing additional levels of security by use of IP addresses and passwords. As NASA and GSFC address issues of homeland security and the dissemination of scientific information, the security mechanisms may branch into biometrics, smart cards, and other advanced mechanisms for security. It will be important for the Archive and the Steering Committee to stay abreast of the security issues that occur.

5.7 Access

The previous life cycle functions are performed for the purpose of ensuring that the bits are available into the future. However, successful preservation must consider changes to access mechanisms, as well as rights management and security requirements over the long term. How do we anticipate what users will need?

5.7.1 Access Mechanisms

Today it is the Web, but there is no way of knowing what it might be tomorrow. It may be possible in the future to enhance the quality of presentation of items from the digital archive based on advances in digitization and browser technologies. Clifford Lynch, Executive Director of the Coalition for Networked Information speaking at a recent CENDI Meeting, suggested that the way to resolve issues of access is to separate the access from the preservation as much as possible. Decisions about standards and practices for preservation should address the "long haul" and they should be relatively stable because of the need to ensure integrity and the cost of revamping, transforming and migrating the preserved content. **However, access mechanisms may differ by the technology and by the needs of the particular audience or community of practice. Access mechanisms may come and go quickly, but this can be easily accommodated if the original design separates the preservation of the content from the expression.**

In the short-term the access to archival materials should be made available via the Library's homepage and if possible in as integrated a way with access to other library resources, both internal and external. However, it should be possible to exclude the archival materials from a search and also to exclude other resources in an archival search.

In addition to the searching of metadata, full text access should be provided to appropriate text objects, when they are available. Not all relevant text objects, particularly legacy items, will have digital versions available, so it is important to include pointers to the physical locations for these items, whether to the GSFC Library or to the records archives. Items determined to be of significant value in an online KM environment, should be considered for digitization.

5.7.2 Rights Management and Security Requirements

One of the most difficult access issues for digital archiving involves rights management. What rights does the archive have? What rights do various user groups have? What rights has the owner retained? How will the access mechanism interact with the archive's metadata to ensure that these rights are managed properly? Management includes providing or restricting access as appropriate, and changing the access rights as the material's copyright and security level changes.

This needs additional work and some consensus across the Center. While flexibility is preferred it must be within a security framework that meets the security regime of NASA, privacy requirements, employee relations requirements, and systems efficiency.

6.0 ADMINISTRATION AND ORGANIZATIONAL INFRASTRUCTURE

In order to provide true archiving that is useful across time and communities, these specific issues must be addressed both by communities of practice and within each local implementation.

Recent work by the CCSDS details the interaction between creators of objects and the archive prior to and during the ingest process. In the case of GSFC, this involves determining the life cycle management for materials of importance to the archive and ultimately to knowledge management. As a natural outgrowth of its mission, the Library will ultimately take physical responsibility for the material. Because of the problems of version control and archive size, it is suggested that the Library take possession of project materials when the project is completed. It would be advisable for a relationship to be developed between the project libraries and the GSFC Library so that in addition to sending materials to the warehouse for potential archiving by NARA, the digital forms would be sent or made available to the Library for spidering and harvesting.

This approach does not solve the problem of access prior to completion of the project. This access could be addressed by making tools available that would allow the project libraries to become OAI compliant. The Open Archive Initiative is intended to create a central metadata repository based on the harvesting of compliant sites. The project libraries could become compliant sites and then the Library could harvest the OAI metadata (which is based on Dublin Core but can accommodate community extensions). This metadata repository would allow searching for project information at any point in the life cycle. Whether a live link would be provided to the actual digital object in the project library would be up to the individual project library manager.

7.0 IMPLEMENTATION PLAN

7.1 Scope of the Implementation Plan

While the scope of digital archiving at GSFC is quite broad, it will be impossible to deal equally with all these objects at the same time. Therefore, this section more narrowly defines the scope and priorities for the follow-up activities. Based on the analysis of the current status of archiving and the interests related to knowledge management and human asset management, the consultant recommends the following priorities. The immediate work should focus first on videos and web pages, since they were successfully piloted during this TO. Because of the importance to knowledge management, project documentation should be the third area of focus. Datasets, including those accessible via the Web, those related to projects and others, will be excluded from the scope. Datasets are extremely large, often require special software to use or visualize, and many of them are successfully managed under the NSDI DACC.

7.1.1 Videos

The Library now encodes, webcasts and archives most of the mini-courses and colloquia series that are held at GSFC based on the successful pilot project. As part of recent initiatives regarding multimedia assets management systems to support knowledge management, the Library provided information regarding human and software/hardware resources needed to institutionalize the capability and to extend it so that other groups can capture the content and then provide it to the Library. (However, since these have not been funded, the costs are included in Appendix B.) Work is underway on speech to text and indexing of the video files. Appropriate metadata is being analyzed, and the team is reviewing how to make this metadata accessible along with metadata from other object types.

However, a large gap exists between the creation of the archival repository for videos and the preservation of those videos. Once the repository infrastructure and access have been completed, the Library will focus its efforts on issues of standards. This is particularly important since full motion video formats are not standardized, and even though many of the standards and tools are currently free, they are provided under the auspices of a particular vendor, such as MicroSoft. In addition, metadata for video preservation and migration strategies to ensure that the videos are available into the future need to be addressed. The work of the Corporation for Public Broadcasting and Warner Brothers should be monitored.

7.1.2 Web Sites

The second major emphasis will be on archiving and preserving web sites. The pilot project provided several valuable lessons. Follow up activities should include: the development of standard workflow procedures for the acquiring of web sites, additional analysis with a broader representation of stakeholder groups surrounding the issues of extent and the archiving of links, and review of the proposed permanence rating system by stakeholders within the GSFC community.

7.1.3 Project Documentation

There is an overlap between the project documentation and the materials available on project web sites. However, it is clear that not all project documentation is published on the open or intranet web. The consultant agrees with the Fraeunhoefer study that web publishing of project documentation should be encouraged. This would allow the library to capture the information in a streamlined fashion using the same techniques used for the capture of other web materials from the center.

7.2 Resources Needed

As part of the Multimedia Asset Management proposal, the Library developed estimates for the costs of human resources, software, hardware, training, implementation and ongoing maintenance. The appropriate estimates from the MAMS Cost Proposal (August 2002) have been reorganized in Appendix B to reflect the major requirements of archiving – development of the repository/depository, spidering, harvesting, and the video encoding, indexing, webcasting and archiving.

7.2.1 Human Resources

As identified in the cost estimate in Appendix B, there are human resources required for archiving on a long-term basis. The consultant suggests that a full time mid-level staff person should be designed as the digital archivist. This person needs a thorough understanding of the issues related to digital preservation and the technologies that can be employed. A librarian with sufficient exposure to these issues and a systems background would qualify. In addition, this person will serve as a collection manager, raising awareness among the GSFC community of the importance of digital preservation and working with the webmasters, project librarians and video capture specialists to ensure that resources of long-term value to the GSFC community are identified, captured, organized and maintained. This person will also be responsible for following standards and monitoring other preservation activities from which GSFC can benefit, within NASA, within the government and in the private sector.

In addition to the digital archivist, support will be needed from a library technician or specialist level to perform some necessary digitization of analog materials. A system administrator will be needed to support the computer hardware/software environment. Periodic support will also be needed from a database management specialist and a programmer.

7.2.2 Software/Hardware

One of the key lessons learned from the web capture pilot was the need for an adequate hardware and software configuration. The web capture mechanisms are currently being tested in an extended production-like pilot; this pilot has already identified the need for storage that is not only adequate for current needs but extensible and well managed into the future. In the next phase of the work, it will be important to determine those archived resources that require online accessibility, those that can be near-line and those that can be stored off-line.

The cost estimate for the repository/depository to support all object types included in Appendix B is based on the following proposed system configuration:

Sun Enterprise 450
Solaris 9 Operating System
Monitor
Video Card
Sun StorEdge D1000 Array
VERITAS Volume Manager
SCSI Host Adapter

Software needs for the spidering and harvesting will be identified as a follow-up to this TO.

7.2.3 Proposed Timeline

Subtask	Start Date	End Date
Operational Plan	Upon receipt of new TO	First draft - 4 months after receipt Final draft – 2 months after first draft Final plan as modified by other TO activities – 2 months after final draft
Framework	Upon receipt of new TO	5 months after receipt
Captured Project Documentation	2 months after receipt of new TO	End of performance period
Captured Web Sites	Upon receipt of new TO	End of performance period
Video Capture/Web Casting	Upon receipt of new TO	End of performance period
Follow Digital Archiving Standards/Best Practices	Upon receipt of new TO	End of performance period

Appendix A: Selection of Project Documentation

Analysis of the Archiving of Product Information Types for Engineers

Generic Information Data Types	**Archive-Do/Don't**
ADP - Acceptance Data Package	
Package for xxx	Don’t Archive
ANYS - Analysis of	
Anomalies, Failures, Risk	Archive
Attitude Control Subsystem Design & Performance	Archive
Thermal Subsystem Design & Performance	Archive
Communications Subsystem Design & Performance	Archive
Electrical Subsystem Design & Performance	Archive
Power Subsystem Design & Performance	Archive
Interfaces	Archive
Interface between xxx & yyy	Archive
Parts Derating	Archive
Sensors & Actuators	Archive
Engineering Study Report	Archive
Worst Case Performance	Archive
On-orbit Mission Operation	Archive
Cost study for xxx	Archive
CM - Configuration Management	
Plans for xxx	Archive
LEGL - Legal	
Memo of understanding for xxx	Don’t Archive
Request of Proposal for xxx	Archive
Statement of work for xxx	Archive
Contract for xxx	Don’t Archive
LOG - Information tracking	
Certification log for xxx	Don’t Archive
Changes	Archive
Documents & Drawings	Archive
Drawing package for xxx	Don’t Archive
Documentation tree for xxx	Archive
Drawing tree for xxx	Archive
Photos & Videos Log	Archive
Functional diagrams of xxx	Archive
MEMO - Memorandums	
Approvals	Don’t Archive
Trip reports	Archive
Technical correspondence	Archive
Anomolies, failures, problems	Archive
General correspondence	Don’t Archive
MGMT - Management	
Schedule/Reschedule for xxx	Don’t Archive

Progress report for xxx	Don’t Archive
Work breakdown structure for xxx	Archive
Status report for xxx	Don’t Archive
Systems and Interface	Archive
Mission presentations	Don’t Archive
Lessons learned	Archive
Problem charts	Archive

MISC - Miscellaneous

Photos	Archive
Videos	Archive
Drawings	Archive
Articles	Archive

PAR - Parts

Material list for xxx	Don’t Archive
Parts list for xxx	Don’t Archive
Wire list for xxx	Don’t Archive
Failure report for xxx	Archive
Malfunction reports for xxx	Archive
Problem report for xxx	Archive
Test failure report for xxx	Archive
Nonstandard parts and material	Archive
Alerts and safe alerts	Don’t Archive

PLAN - Plans

Implementation plan for xxx	Archive
Development plan for xxx	Archive
Assembly plan for xxx	Don’t Archive
Handling and shipping plan for xxx	Don’t Archive
Strength verification plan for xxx	Archive
Flight operations plan	Archive
Parts control plan	Don't Archive
Performance assurance plan for xxx	Archive
Quality assurance plan for xxx	Archive
Qualification and acceptance plan for xxx	Archive
Software management plan	Archive

PROC - Procedures

General procedures for xxx	Archive
Calibration procedures for xxx	Don’t Archive
Protoflight test procedures	Don’t Archive
Functional test procedures for xxx	Archive
Acceptance test procedures for xxx	Archive
Integration procedures for xxx	Archive
Verification procedures for xxx	Archive
Test procedures for xxx	Archive
Mission procedures for xxx	Archive
Procedures for modification of xxx	Don’t Archive
Assembly procedures for xxx	Don’t Archive
Recertification procedures for xxx	Don’t Archive
Installation procedures for xxx	Don’t Archive
Launch procedures	Don’t Archive

Safe to Mate procedures	Don’t Archive
REVW - Reviews	
Concept review	Archive
Functional assessment	Archive
Preliminary design review for xxx	Archive
Critical design review for xxx	Archive
Failure review for xxx	Archive
Design review of xxx	Archive
Action items and responses for xxx	Archive
Peer review	Archive
Non-advocate review	Don’t Archive
Configuration review for xxx	Archive
Pre-ship review for xxx	Don’t Archive
Flight readiness review	Archive
Software for xxx review	Archive
Calibration review for xxx	Archive
RUN - Run test	
Run test procedure for xxx	Archive
Verification and/or Validation test procedure for xxx	Archive
Integration & test procedure for xxx	Archive
Electrical integration for xxx	Archive
SDL - Software delivery letter	
Source code	Archive
Delivery description	Archive
SPEC - Specifications and Requirements	
Specifications for xxx	Archive
Performance Assurance Requirements	Archive
Interface control document for xxx	Archive
Handling requirements	Don’t Archive
Workmanship requirements for xxx	Don’t Archive
Algorithm document for xxx	Archive
Flight software requirements	Archive
Environment Specifications	Archive
Science requirements	Archive
Functional description and requirements	Archive
TEV - Test and Verification	
Test objective for xxx	Archive
Thermal test plan	Archive
Vibration test plan	Archive
Mechanical subsystem test plan	Archive
Software test validation/verification plan	Archive
Protoflight test plan for xxx	Archive
Component test plans for xxx	Archive
Performance test plan for xxx	Archive
Verification & validation test plan for xxx	Archive
Integration test plan for xxx	Archive
Interface test plan of xxx	Archive
Qualification and Acceptance testing of xxx	Archive

TR - Test report

Test report for part xxx	Archive
Test report for interface xxx	Archive
Test report for subsystem xxx	Archive
Test report for integration of xxx	Don’t Archive
Test report for software	Archive

APPENDIX B: RESOURCE ESTIMATE FOR DIGITAL ARCHIVING

COST ESTIMATE FOR DIGITAL ARCHIVING

		Year 1	Year 2	Year 3	Year 4	Year 5	Year 6
Repository/ Depository	Database & Retrieval	45,000	0	0	0	0	0
	Training (1)(3)(4)	3,000	1,500	1,575	1,654	1,736	1,823
	Maintenance Fee (6)	9,000	9,000	9,000	9,000	9,000	9,000
	Hardware (5)	49,154	1,600	1,600	1,600	1,600	1,600
	FTEs (2) (4)	200,000	208,000	216,320	224,973	233,972	243,331
	TOTAL	**$306,154**	**$220,100**	**$228,495**	**$237,227**	**$246,308**	**$255,754**
Spider	Software	3,000	0	0	0	0	0
	Maintenance Fee (6)	0	600	600	600	600	600
	TOTAL	**3,000**	**600**	**600**	**600**	**600**	**600**
Harvester	Programming	60,000	6,000	6,000	6,000	6,000	6,000
	FTEs (2) (4)	97,500	101,400	105,456	109,674	114,061	118,624
	TOTAL	**$157,500**	**$107,400**	**$111,456**	**$115,674**	**$120,061**	**$124,624**
Video Capture	Maintenance Fee (6)		3,000	3,000	6,000	6,000	6,000
	Hardware (7)	30,000		30,000			30,000
	FTEs (2) (4)	45,000	45,000	45,000	45,000	45,000	45,000
	TOTAL	**$75,000**	**$48,000**	**$78,000**	**$51,000**	**$51,000**	**$81,000**

(1) Training is defined as direct payments to vendors. It does not include the time employees or contractors spent being trained.

(2) FTEs are contractor staff only.

(3) Training includes training for main software and database and retrieval software.

(4) Numbers are adjusted for inflation: training at 5% per year increase and FTEs at 4% per year increase.

(5) Outyear cost for hardware is the yearly warranty of disk array.

(6) Based on maintenance fee of 20% of purchase price.

(7) Assumes purchase of 3 portable video capture set ups over the 6-year period.